Use of Condition Numbers for Shortcut Experimental Design

C. R. Kaplan
Combustion and Fuels Branch
Naval Research Laboratory
Washington, DC 20375

J. W. Gentry
Department of Chemical
and Nuclear Engineering
University of Maryland
College Park, MD 20742

Introduction and Background

During the mid-1940's, several mathematical criteria were developed to determine the effect of round-off error in the solution of large sets of linear algebraic equations with digital computers (von Neumann and Goldstine, 1947; Turing, 1948). Recent developments of the use of condition numbers have shown that they can be used as a criterion for the interpretation of physical measurements of airborne particulate matter. Farzanah et al. (1985) have used them as a criterion for the selection of indicator elements used in determining the relative importance of source strengths in source receptor models. Cooper (1975) has used condition numbers to indicate the magnification of error when inverting diffusion battery data and impactor data (Cooper and Spielman, 1976). They have been used by Yu (1983) to provide a means of comparison of different size classifiers and for the evaluation of inversion algorithms in determining the size distribution of aerosols. And most recently, they have been used as a criterion for the choice of analysis of multimodal distributions in both atmospheric aerosol measurements and droplet breakup studies (Kaplan et al., 1985).

Further development has led to the use of condition numbers as a measure of error sensitivity, and as a criterion for data evaluation and experimental design. As used here, the condition number represents the maximum amount by which a perturbation in an experimental measurement will be transmitted to the unknown variables. When the condition number is a minimum, small errors in the experimental measurements have the least effect on the unknown variables. Our criterion for choosing experimental conditions is based on minimization of the condition number, which is calculated from the matrix that is formed when using a least-squares regression technique to fit the data to a linearized form of an n-dimensional power law correlation. Many of the empirical correlations used in the estimation of heat and mass transfer coefficients, as well as in many other fields of chemical engineering, are of the power law type.

Recent work in this area (Kaplan, 1985) has shown that the distribution of the data points affects the accuracy of fitting the data to a power law correlation; broader distributions (less interdependence between the independent variables) give more accurate solutions. Also, condition numbers indicate that experimental data will fit a two-parameter correlation more accurately than those of higher order (Kaplan, 1985), unless the multidimensional data are very broadly distributed, hence emphasizing the need to reduce the proliferation of unnecessary terms in empirical correlations. This kind of result has been confirmed by Cooper (1984), who showed that nucleate boiling coefficients can be described better by a two-parameter correlation in reduced pressure than by a multiparameter correlation in terms of several dimensionless heat transfer groups.

The utility of the method outlined below lies in that it provides *a priori* criteria for the degree of interdependence of the independent variables. For example, in heat transfer applications the Nusselt number is frequently correlated with the Reynolds and Prandtl numbers. These two variables are nominally independent, but if all the measurements were made along paths of increasing temperature, and if the variations in Reynolds and Prandtl numbers were primarily due to the temperature variation, the dimensional groups would effectively be interdependent. Correlations based on such data would be ambiguous. In cases where the independent variables are two, it is conceivable that engineering intuition would suffice. However, when there

Correspondence concerning this paper should be addressed to C. R. Kaplan.

are more than four such groups, more rigorous criteria are necessary. Furthermore, it is not sufficient that these criteria merely indicate dependence, but that they should quantify the degree of interdependence. The criteria specified below were directed toward this goal.

Definition of Condition Numbers

Consider the set of N simultaneous linear algebraic equations, written as

$$[A][X] = [B] \tag{1}$$

where $[A]$ corresponds to the $N \times N$ matrix of known coefficients that are determined from theory or empirical correlations; $[X]$ is the $N \times 1$ vector of unknown variables; and $[B]$ is the $N \times 1$ vector corresponding to experimental measurements. Mathematically, condition numbers are criteria that relate the changes in the unknown variables $[X]$ to an error in the experimental measurements $[B]$. If small errors in the experimental measurements, B_i, have an insignificant effect on the unknown variables, X_j, then the system is considered to be well-conditioned. If however, the effect is large, then the system is considered to be ill-conditioned.

The condition number is defined as

$$C(N, P) = \|A\|_P \|A^{-1}\|_P \tag{2}$$

where $\| \ \|_P$ corresponds to the pth matrix norm. The three matrix norms (Noble, 1969), most commonly used are:

1. $\|A\|_1 = \max_j \Sigma_i |a_{ij}|$, which represents the maximum value of the sum of the absolute values of the elements in each column.
2. $\|A\|_2 = \{$maximum eigenvalue of $[A]\}$ for a symmetric matrix only.
3. $\|A\|_\infty = \max_i \Sigma_j |a_{ij}|$, which represents the maximum value of the sum of the absolute values of the elements in each row.

The condition numbers associated with each of the norms are:

$$\begin{aligned} C(N, 1) &= \|A\|_1 \, \|A^{-1}\|_1 \\ C(N, \infty) &= \|A\|_\infty \, \|A^{-1}\|_\infty \\ C(N, 2) &= \|A\|_2 \, \|A^{-1}\|_2 \end{aligned} \tag{3}$$

It is clear from this definition that the condition number depends only on the matrix of coefficients $[A]$, and is independent of the measurements vector $[B]$. Hence, it is reasonable to suggest that condition numbers may serve as criteria for experimental design. Second, the condition number depends both on $[A]$ and its inverse. For a set of equations that are nearly singular, the coefficients of the inverse matrix are large; therefore the condition number is large and the system is considered to be ill-conditioned.

The condition number has two important properties. The first one

$$\frac{\|\Delta X\|_P}{\|X\|_P} \le C(N, P) \frac{\|\Delta B\|_P}{\|B\|_P} \tag{4}$$

implies that the relative error in the norm of the solution is bounded by the product of the condition number and the relative error in the norm of the measured variables. Hence, the condition number represents the maximum amount by which any perturbation in the measurement will be transmitted to the solution, and is clearly a measure of error sensitivity. Second, from the norm equivalence theorem (Ortega and Rheinboldt, 1970) for any two norms, p_1 and p_2,

$$a_1 C(N, P_2) \le C(N, P_1) \le a_2 C(N, P_2) \tag{5}$$

where $a_2 > a_1 > 0$. Typical values are $a_1 \sim 1/2$, and $a_2 \sim 2$. This inequality implies that if a set of equations is ill-conditioned, the condition numbers for any matrix norm will be large. Consequently, to test the system of equations it is necessary to calculate only one norm, preferably the one requiring the least computation.

To some extent, condition numbers can be affected by normalization of $[A]$, the matrix of coefficients. If the rows are normalized such that the sum of the absolute value of the elements is one, called uniform row normalization, then $C(N, \infty)$ is minimized. This does not imply that $C(N, \infty)$ is less than $C(N, 2)$, or that for some other normalization an even lower condition number can be obtained. The condition numbers that are simplest to use are $C(N, 1)$ and $C(N, \infty)$. To make the condition numbers as sensitive and meaningful as possible, we suggest that the elements of $[A]$ be scaled by uniform row normalization when $C(N, 1)$ or $C(N, \infty)$ is calcualted. By doing so, a large condition number will be due to a singularity in $[A]$, rather than due to the fact that the elements in $[A]$ are large numbers; this enables one to make direct comparisons between condition numbers of different matrices. Of course, direct comparisons can only be made of condition numbers calculated from the same type of matrix norm using the same normalization technique.

Applications

Let us assume that we have n data points, x_i, y_i vs. g_i for $i = 1, 2, \ldots, n$, and we wish to determine the coefficients a, b, and c of the correlation

$$g = a x^b y^c \tag{6}$$

To calculate the condition number, we must be able to represent the correlation in terms of x_i, y_i, g_i in matrix form, as a set of simultaneous linear algebraic equations. Hence, we linearize the power law correlation by taking the logarithm of both sides of Eq. 6, and then solve the following set of equations to determine the unknown coefficients a, b, and c.

$$\begin{array}{ccccc} [A] & & [X] & = & [B] \\ \begin{bmatrix} 1 & \ln x_1 & \ln y_1 \\ 1 & \ln x_2 & \ln y_2 \\ \vdots & \vdots & \vdots \\ 1 & \ln x_n & \ln y_n \end{bmatrix} & & \begin{bmatrix} \ln a \\ b \\ c \end{bmatrix} & = & \begin{bmatrix} \ln g_1 \\ \ln g_2 \\ \vdots \\ \ln g_n \end{bmatrix} \end{array} \tag{7}$$

If one has as many data points as unknowns, (i.e., $n = 3$ in this case), then we can solve the problem deterministically, and would calculate condition numbers according to Eq. 2. However, in most cases we have many more data points than unknowns, and would use a least-squares regression technique to solve for the unknown coefficients. Minimizing the sum of the squares, the least-squares regression solution is

$$\begin{bmatrix} n & \Sigma \ln x_i & \Sigma \ln y_i \\ \Sigma \ln x_i & \Sigma(\ln x_i)^2 & \Sigma \ln x_i \ln y_i \\ \Sigma \ln y_i & \Sigma \ln y_i \ln x_i & \Sigma(\ln y_i)^2 \end{bmatrix} \begin{bmatrix} \ln a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \Sigma \ln g_i \\ \Sigma \ln g_i \ln x_i \\ \Sigma \ln g_i \ln y_i \end{bmatrix} \qquad (8)$$

where the summations are taken from 1 to n. This corresponds to (Lawson and Hanson, 1974)

$$[A]^T[A][X] = [A]^T[B] \qquad (9)$$

The condition number for the least-squares regression is given by

$$C(N, P) = \|(A^TA)\|_P\|(A^TA)^{-1}\|_P \qquad (10)$$

Minimization of this condition number will provide a good criterion for choosing the optimal experimental parameters (values of the independent variables x_i and y_i) to yield the most accurate values of the coefficients a, b, and c. It should be noted here that a conventional criterion for optimal design for linear regression problems has been treated in the statistical literature (Steinberg and Hunter, 1984) and is based on minimization of the determinant of $(A^TA)^{-1}$. This criterion is essentially equivalent to the one proposed in this paper, as when the determinant of $(A^TA)^{-1}$ is minimized, the condition number of (A^TA), defined in Eq. 10, is also minimized.

The condition number for the least-squares regression technique is always greater than that defined in Eq. 2 for a deterministic solution. However, by using an overspecified system (more data points than unknowns), one minimizes the relative error in the measurement $\|\Delta B\|_P/\|B\|_P$. According to Eq. 4, to minimize the relative error in the solution, $\|\Delta X\|_P/\|X\|_P$, there is a tradeoff between reduced measurement error and larger condition number. We have found that the condition number increases slowly with the addition of more measurements, such that the relative error in the solution is lower with least-squares regression than with a deterministic solution. By suitably choosing the location of experiments, we can minimize the condition number for the least-squares regression solution, and hence minimize the relative error in the solution.

As an example, let us assume that the raw data, $\bar{x}$ and $\bar{y}$, are distributed according to

$$\bar{x}^2 + \frac{\bar{y}^2}{q^2} = t^2 \qquad (11)$$

For $q = 1$, the data are located on a circle; for $q \neq 1$, the data are located on an ellipse. Letting $t^2 = 1$, the circle or ellipse is centered around the origin with the major axis coincident with the x axis. Since the condition number is calculated from the matrix $[A^TA]$, whose elements represent summations of the independent variables, Eq. 8, we must move the ellipse from being centered around the origin, as this would result in the summations being zero. Hence, we rotate the circle or ellipse 45 degrees in the counterclockwise direction such that the major axis is coincident with the diagonal line $x = y$. The circle or ellipse was rotated by multiplying the $[x, y]$ vector (for each data point) by a rotation matrix, $[R]$,

$$\begin{bmatrix} x \\ y \end{bmatrix} = [R] \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \qquad (12)$$

where

$$[R] = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

The rotation matrix was constructed such that the rows and columns are orthogonal and orthonormal, and satisfies the requirement that $[R]^T[R] = [I]$. The criteria for construction of a generalized rotation matrix is explained in more detail in the Supplementary Material. The eight data points distributed on the rotated circle or ellipse were then fitted to a two-dimensional power law correlation using a least-squares regression technique and the condition number calculated, using the one ($p = 1$) norm, according to Eq. 10. The results are shown in Figure 1 for different values of q. As the dimension of the minor axis is decreased, (i.e., as q is decreased), the condition number increases. This suggests that as the data are more broadly distributed, as indicated by increasing the axes of the ellipsoid, there is less interdependence among the independent variables, the condition number decreases, and the data fit the power law correlation more accurately.

Similar results, using the one ($p = 1$) norm, are shown in Table 1 for the analogous case of fitting data to a three-parameter correlation

$$g = ax^by^cz^d \qquad (13)$$

in which the 14 data points are distributed on prolate and oblate spheroids,

$$\bar{x}^2 + \frac{\bar{y}^2}{q_1^2} + \frac{\bar{z}^2}{q_2^2} = 1 \qquad (14)$$

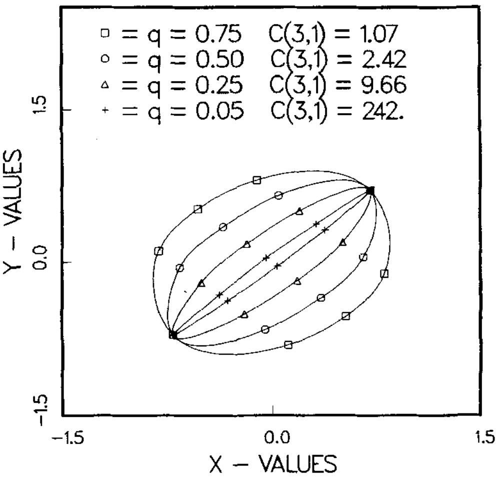

Figure 1. Condition numbers to fit data points from Eq. 11 to a two-dimensional correlation, Eq. 6.

Table 1. Condition Numbers to Fit Data Points, from Eq. 14, to a Three-Parameter Power Law Correlation, Eq. 13

q_1	q_2	$C(4,1)$
	Prolate Spheroids	
1	1	1.34
0.5	0.5	3.97
0.25	0.25	1.69×10^1
0.1	0.1	1.08×10^2
0.05	0.05	4.31×10^2
0.01	0.01	1.07×10^4
	Oblate Spheroids	
1	2	8.29
1	3	2.07×10^1
1	25	1.56×10^3

The results from Figure 1 and Table 1 suggest that there is an envelope region, bounded by the ellipse or spheroid (ellipsoid), outside of which the condition number is small. As the data points are placed inside this envelope region, the condition number increases. We propose that for purposes of experimental design it is beneficial to predetermine the location of the envelope as a bounding region for taking measurements, such that the condition number is small. In constructing this bounding region it is not only important to minimize the condition number, but also the product of the condition number and the relative error in the norm of the measured variables, as expressed in Eq. 4. As the data points are more broadly distributed, the condition number decreases; however, the relative experimental error may possibly increase, such that the relative error in the norm of the solution is not a minimum.

Data Evaluation

We illustrate a procedure for evaluating a set of well-established heat transfer data on three oils (Seider and Tate, 1936) to estimate the effectiveness of a correlation. We used 67 data points of Reynolds number, varying from 3 to 2,110, and Prandtl number, varying from 151 to 16,700. The data were normalized using a logarithmic transformation such that all values of the normalized variables lie between 0 and 1. The data were rotated and translated such that the elliptical pattern lies along the x axis and is centered around the origin, as shown by the data points in Figure 2. It is recommended that the data points be centered in this way to simplify construction of the bounding ellipse.

One may then calculate the equation of the bounding ellipse from the major and minor axes of the elliptical pattern formed by the outer boundary of the data points. The major axis, as estimated from the boundary of the data points in Figure 2, is 0.703. Ellipses with a major axis of $t = 0.703$ and varying minor axes, q, as calculated according to Eq. 11, are shown in Figure 2. The ellipse with the minor axis of $q = 0.15$ encompasses most of the data points, while that with $q = 0.3$ encompasses all of the data points.

Condition numbers for the different bounding ellipses, each composed of 12 data points, are shown in Figure 2. Clearly, the condition number decreases as the minor axis of the ellipse increases, and hence the level of interdependence between normalized Prandtl and Reynolds numbers decreases. For the spread of data points located on the bounding ellipse at which the condition number is sufficiently low, one would then back-calculate corresponding values of the Reynolds and Prandtl number by reversing the normalization and rotation procedures. These correspond to the values of Reynolds and Prandtl numbers at which one should take experimental measurements.

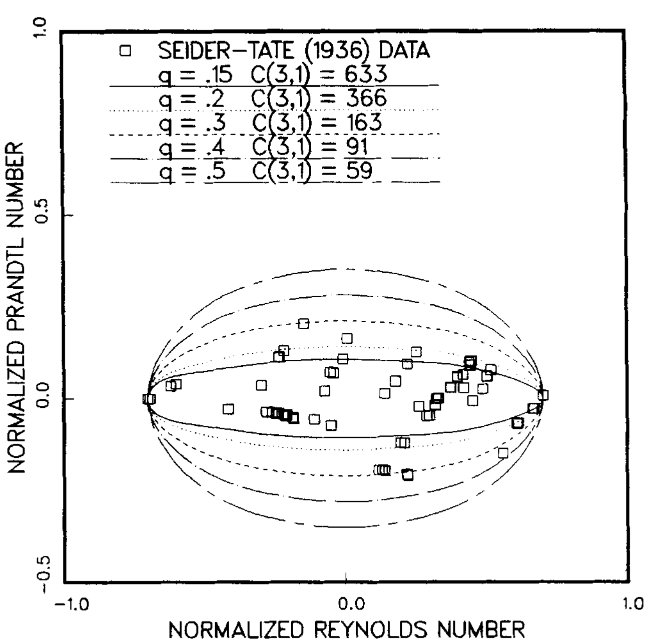

Figure 2. Condition numbers for different bounding ellipses.

The decision as to what constitutes a sufficiently low condition number is somewhat arbitrary, and depends on the particular experimental circumstances. It is fairly safe to say that the data will not accurately fit a power law correlation when the condition number is on the order of 10^5 or higher; however, for condition numbers of 10^3–10^4, more judgment is necessary. We propose that not only the condition number be minimized, but also that one should minimize the product of the condition number and the relative error in the norm of the measured variables, $\|\Delta B\|/\|B\|$, as specified in Eq. 4.

The relative error in the norm of the measured variables is dependent on the particular experimental conditions. The condition number decreases as one widens the distribution of the independent variables; however, it may be difficult or impossible to make accurate measurements at the far-reaching regions, such that the relative error in the measured variables is high. Clearly, there is a trade-off between minimization of the condition number and experimental error in this case. Hence, based on the particular experimental conditions, one must estimate the experimental error and make a trade-off between minimizing it and the condition number.

In summary, we have proposed that the condition number may be used as a criterion for experimental design. The condition number is used as a criterion for construction of a bounding region to quantify the degree of interdependence among independent variables. This method would be most valuable for predetermining experimental conditions such that experimental parameters may be chosen to result in accurate and meaningful empirical power law correlations.

Notation

a, b, c, d = coefficients in power law correlations
$[A]$ = matrix of known coefficients
$[B]$ = vector of experimental measurements
$C(N, P)$ = condition number based on pth matrix norm
g = dependent variable in correlations
$[I]$ = identity matrix
n = number of data points (measurements)
N = number of simultaneous linear algebraic equations
q = minor axis parameters for equations of ellipses, spheroids
$[R]$ = rotation matrix
t = radius term in equations of ellipses, spheroids
x, y, z = independent variables in correlations
$[X]$ = vector of unknown variables
$\| \ \|_p$ = pth matrix norm

Literature cited

Cooper, D. W., "Diffusion Battery Data Inversion," *Proc. 68th Ann. Meet. Air Pollution Control Ass.,* Boston (1975).

Cooper, D. W., and L. A. Spielman, "Data Inversion Using Nonlinear Programming with Physical Constraints: Aerosol Size Distribution Measurement by Impactors," *Atm. Environ.,* **10,** 723 (1976).

Cooper, M. G., "Heat Flow Rates in Saturated Nucleate Pool Boiling—A Wide-Ranging Examination Using Reduced Properties," *Adv. Heat Trans.,* **16,** 157 (1984).

Farzanah, F. F., C. R. Kaplan, P. Y. Yu, J. Hong, and J. W. Gentry, "Condition Numbers as Criteria for Evaluation of Atmospheric Aerosol Measurement Techniques," *Environ. Sci. Tech.,* **19,** 121 (1985).

Kaplan, C. R., "Systematic Use of Condition Numbers in Experimental Design," M.Sc. Paper, Univ. Maryland (1985).

Kaplan, C. R., R. V. Calabrese, and J. W. Gentry, "Condition Numbers: Application to Correlations, Distribution Analysis, and Experiment Location," *Proc. 1985 Ann. Meet. Am. Ass. Aerosol Research,* Albuquerque, NM, (Nov., 1985).

Lawson, G. L., and R. J. Hanson, *Solving Least-Squares Problems,* Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ (1974).

Noble, B., *Applied Linear Algebra,* Prentice-Hall, Englewood Cliffs, NJ (1969).

Ortega, J. M., and W. C. Rheinboldt, *Iterative Solutions of Nonlinear Equations in Several Variables,* Academic Press, New York (1970).

Seider, E. N., and G. E. Tate, "Heat Transfer and Pressure Drop of Liquid in Tubes," *Ind. Eng. Chem.,* **28,** 1429 (1936).

Steinberg, D. M., and W. G. Hunter, "Experimental Design: Review and Comment," *Technometrics,* **26,** 71 (1984).

Turing, A. M., "Rounding-off Errors in Matrix Processes," *Q. J. Mech. Appl. Math.,* **1,** 287 (1948).

von Neumann, J., and H. Goldstine, "Numerical Inverting of Matrices of Higher Order," *Bull. Am. Math. Soc.,* **53,** 1021 (1947).

Yu, P. Y., "The Simulation and Experiment for Determination of Size and Charge Distributions of Fibrous Particles from Penetration Measurements," Ph.D. Diss., Univ. Maryland (1983).

Manuscript received Feb. 3, 1986, and revision received July 22, 1986.

See NAPS document no. 04445 for 4 pages of supplementary material. Order from NAPS c/o Microfiche Publications, P.O. Box 3513, Grand Central Station, New York, NY 10163. Remit in advance in U.S. funds only $7.75 for photocopies or $4.00 for microfiche. Outside the U.S. and Canada, add postage of $4.50 for the first 20 pages and $1.00 for each of 10 pages of material thereafter, $1.50 for microfiche postage.